

# SYSTEM AND METHOD FOR ROUTING NETWORK TRAFFIC THROUGH WEIGHTED ZONES

5

**Inventors:** Jeremy N. Shapiro & Stephen A. Jay

## BACKGROUND

10        This invention relates to the field of computer systems and networking.  
More particularly, a system and methods are provided for routing network traffic  
through fault zones.

Routing of packets and/or other electronic communications through a  
network depends upon accurate routing. Network nodes that perform routing  
15 (e.g., switches, routers) usually maintain routing tables indicating how to route a  
communication addressed to a particular node.

In order to improve performance of the network, multiple paths to a given  
destination node may be available. The path traveled by a communication to the  
destination node may depend upon factors such as network congestion, which  
20 network links are functional or nonfunctional, the cost of traversing a particular  
path, and/or other factors. However, existing networks do not make efficient  
routing decisions based on how many current paths traverse particular network  
components or links (e.g., switches, wireless links). In particular, existing  
networks typically do not consider or promote the distribution of paths among  
25 different fault zones.

Also, in many network architectures (e.g., TCP/IP networks such as the  
Internet), the construction of routing tables is distributed among multiple nodes.

Because of this distribution of labor, routing algorithms must be relatively simple so that each node is capable of performing them without impacting network traffic.

5

## SUMMARY

Therefore, what is needed is a centralized system and method for evaluating and selecting different network paths, in which the distribution of paths among different fault zones may be considered.

10 In one embodiment of the invention, a system and methods are provided for configuring routing between nodes in a network or subnet. An end node may be associated with multiple identifiers for routing purposes, and therefore multiple paths may exist between two end nodes. Network nodes and components (e.g., switches) are grouped into fault zones. Each physical enclosure of network entities may comprise a separate fault zone. For each zone through which a path  
15 between two nodes passes, a weight is calculated equal to the number of paths between the nodes that traverse that zone.

Path weights are calculated for each path between the nodes, equal to the sum of the weights of each zone in the path. To improve network fault tolerance, new paths may be designed to avoid fault zones and existing paths with high  
20 weights. Instead of fault zones, other criteria may be used to assign weights, such as mean time between failures (MTBF).

## DESCRIPTION OF THE FIGURES

FIG. 1 is a block diagram depicting a subnet with a subnet manager, in  
25 accordance with an embodiment of the present invention.

FIGs. 2A-2B depict the calculation of fault zone weights for different paths between end nodes, in accordance with an embodiment of the invention.

FIG. 3 is a flowchart demonstrating a method of determining routing between nodes in a subnet, according to one embodiment of the invention.

### DETAILED DESCRIPTION

5           The following description is presented to enable any person skilled in the art to make and use the invention, and is provided in the context of particular applications of the invention and their requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art and the general principles defined herein may be applied to other embodiments and appli-  
10   cations without departing from the scope of the present invention. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

          The program environment in which a present embodiment of the invention  
15   is executed illustratively incorporates a general-purpose computer, a special-purpose computer or a network component (e.g., a switch, a network interface device). Details of such devices (e.g., processor, memory, data storage, display) may be omitted for the sake of clarity.

          It should also be understood that the techniques of the present invention  
20   may be implemented using a variety of technologies. For example, the methods described herein may be implemented in software executing on a computer system, or implemented in hardware utilizing either a combination of microprocessors or other specially designed application specific integrated circuits, programmable logic devices, or various combinations thereof. In  
25   particular, the methods described herein may be implemented by a series of computer-executable instructions residing on a suitable computer-readable medium. Suitable computer-readable media may include volatile (e.g., RAM)

and/or non-volatile (e.g., ROM, disk) memory, carrier waves and transmission media (e.g., copper wire, coaxial cable, fiber optic media). Exemplary carrier waves may take the form of electrical, electromagnetic or optical signals conveying digital data streams along a local network, a publicly accessible  
5 network such as the Internet or some other communication link.

In one embodiment of the invention, a system and method are provided for configuring, evaluating or selecting among different network paths to a particular node, in which the distribution of paths based on various criteria may be considered. For example, it may be desirable to promote a distribution of paths  
10 among different network fault zones, to improve fault tolerance. Or, it may be desirable to promote a distribution of paths based on the mean time between failure (MTBF) of network entities (e.g., switches, network links).

Embodiments of the invention are described below as they may be implemented within an InfiniBand environment. One skilled in the art will  
15 appreciate, however, that the invention is not limited to any particular network or communication technologies, and may be adapted for various such technologies.

InfiniBand is a communication architecture designed to provide high-speed interconnection between end nodes (e.g., channel adapters) and switches. Within a subnetwork or subnet of InfiniBand nodes, a subnet manager is  
20 responsible for detecting changes to the subnet configuration and updating nodes' routing tables.

FIG. 1 is a block diagram of an InfiniBand subnet, according to one embodiment of the invention. In this embodiment, subnet 102 includes nodes 104, which may be switches or other network entities capable of routing  
25 communications based on a routing table, routing tree or other schedule. Nodes 104 are configured to route communications within subnet 102 as well as the larger network that includes subnet 102.

Subnet 102 also includes end nodes 106, which may be channel adapters (e.g., target channel adapters and/or host channel adapters). End nodes, such as end nodes 106a – 106c may be coupled to clients such as servers, desktops, portable or other types of computers or processing devices.

5           Within subnet 102, one client or computing device coupled to a node or end node is configured to execute a subnet manager. In FIG. 1, subnet manager 110 operates on client 106c. In another embodiment of the invention, the subnet manager may execute on a computer or computing device coupled to a node 104 instead of an end node. For redundancy and fault tolerance, a secondary subnet  
10       manager may be configured to execute on another node of subnet 102 in the event subnet manager 110, or client 106c or the end node coupling client 106c to the subnet, fails.

          Subnet manager 110 is configured to detect changes in the configuration of subnet 102 and to update one or more routing tables for routing communications  
15       within the subnet. After updating a routing table, the subnet manager may disseminate it to all nodes in the subnet that are configured to route communications. One difference between the network environment in FIG. 1 and many traditional networks (e.g., networks employing the Internet Protocol), is that subnet routing decisions are centralized (i.e., in subnet manager 110) rather than  
20       being distributed among multiple nodes.

          In InfiniBand, a single node may have multiple local identifiers (LIDs) for routing. Because it may have multiple LIDs, multiple paths can be defined to that node. Multiple paths may be defined to promote load balancing among the paths or constituent network links, to promote fault tolerance, and/or for other purposes.  
25       However, for each LID of a node, only one path will be active at a time.

          In one embodiment of the invention, a method is provided for comparing multiple paths to a node having multiple LIDS, and/or for choosing between

multiple paths based on one or more selected criteria (e.g., fault tolerance, mean time between failure (MTBF)).

To promote fault tolerance, it is desirable to define paths through a subnet such that failure of one fault zone affects a minimal number of paths and will not  
5 eliminate all paths to a node. MTBF may be considered in order to favor network links or components that are less likely to fail.

In an embodiment of the invention in which paths are defined to promote fault tolerance, multiple fault zones may be defined in a single network or subnet. Each zone can contain any subset of the nodes and end nodes in the network or  
10 subnet, and fault zones may overlap. However, in one implementation, all nodes and/or end nodes within a single physical enclosure are part of the same fault zone.

In this embodiment, for each LID of an end node (or other destination) in its subnet, a subnet manager notes the fault zones that are traversed by the current  
15 path to that LID. For each fault zone that is traversed by one or more paths to the end node, a weight is defined that is equal to the number of paths to the end node that traverse that fault zone.

FIG. 2A demonstrates how this weighting process may be applied, according to one embodiment of the invention. In FIG. 2A, end node 202 has LID  
20 S, while end node 204 has two LIDs: X and Y. Thus, there are two paths from end node 202 to end node 204, one using a destination LID of X and the other using a destination LID of Y.

Both paths traverse fault zone A, but the path to LID X traverses fault zone C while the path to LID Y traverses fault zone D. Thus, fault zone A receives a  
25 zone weight, or path count, of two, and fault zones C and D both receive zone weights of one.

A path weight can then be defined to characterize the entire path to a LID. In one embodiment, a path weight is equal to the sums of the zone weights or path counts for each fault zone in the path. Thus, in FIG. 2A, the path weights of the two paths are calculated as:

5                    path S→X: zone weight(A) + zone weight(C) = 2 + 1 = 3  
                     path S→Y: zone weight(A) + zone weight(D) = 2 + 1 = 3

In one embodiment of the invention, when a new path becomes available between two end nodes in a subnet, the subnet manager calculates the path weight for the new path, and determines if it would be better to use the new path instead  
10 of the current path.

FIG. 2B depicts the subnet of FIG. 2A with a new path available from LID S to LID Y. Thus, the subnet manager will determine whether the new path, through fault zones B and D, is preferable to the current path through A and D. If the new path were adopted, the zone weights of fault zones B and D would both  
15 be one, and the path weight would be:

$$\text{path S} \rightarrow \text{Y: zone weight(B) + zone weight(D) = 1 + 1 = 2}$$

Thus, the new path from LID S to LID Y would be adopted. Note that this would cause the path weight of path S→X to improve, because the zone weight of fault zone A would drop to one.

20            When a new path is selected, the subnet manager may update a routing tree or table, and disseminate the routing information to nodes in the subnet.

In the embodiment of the invention depicted in FIGs. 2A-2B, as more paths are defined between a pair of nodes in a subnet, the fault zones with the most paths will be weighted higher than other fault zones. Thus, new paths will  
25 tend to avoid the over-subscribed zones.

In one alternative embodiment of the invention, the calculation of zone weights may consider all system paths through a fault zone, not just all paths

between a specified pair of end nodes. In this alternative embodiment, the weight assigned to a zone is proportional to the number of system paths that traverse that zone.

To address network issues or criteria other than fault zone reliability, weights may be assigned or calculated on some other basis. For example, weights could be based on the mean time between failure (MTBF) of a set of network entities (e.g., switches, links). In this example, entities that fail relatively frequently will receive worse weights, and hence fewer paths will traverse them. As one result, if such an entity does fail, it will likely have a less drastic effect on network traffic than it would otherwise. An entity's weighting may be proportional or inversely proportional to its expected MTBF.

Other factors or criteria that may be used to assign weights may include link speed, hop count, quality of service (QoS), financial cost, etc. Calculations of weights and paths may use any combination of these and/or other factors.

FIG. 3 is a flowchart demonstrating a method of calculating fault zone weights and path weights, according to one embodiment of the invention.

In operation 302, a network or subnet is configured with multiple fault zones. Each fault zone includes any number of network switching nodes (e.g., switches, routers), end nodes (e.g., channel adapters, network interface adapters), clients (e.g., servers, input/output subsystems, workstations) and/or other devices.

In operation 304, a subnet manager is installed and configured on a client within the subnet. The subnet manager learns the configuration of the subnet, to include nodes and their identifiers, links between nodes, etc.

In operation 306, the subnet manager identifies one or more paths from one node to an end node. The end node may have multiple identifiers (e.g., LIDs, network addresses).



In operation 308, for each fault zone that is used by at least one of the paths to the end node, the subnet manager calculates a zone weight equal to the number of paths to the end node that traverse that zone.

5 In operation 310, a path weight is calculated for each path, and is equal to the sums of the zone weights for each fault zone used by the path. The zone and/or path weights may be stored by the subnet manager for use in generating a routing table or tree for the subnet.

10 In operation 312, it is determined whether any more paths in the subnet need to be examined and given weights. The subnet manager may examine all paths between each pair of nodes within the subnet or just a subset of all paths. If additional paths need to be examined, the method returns to operation 306.

15 In operation 314, the subnet manager assembles and disseminates routing information to appropriate network entities (e.g., switches, routers). Illustratively, the subnet manager selects one path, based on path weights, between a given pair of nodes or node identifiers. Thus, each network switching element will apply the specified routing for that path until changed by the subnet manager.

In operation 316, the subnet manager learns of a new path to an end node or one identifier of an end node, for which a different path is currently specified in the subnet routing information.

20 In operation 318, the subnet manager calculates zone and path weights for the path, as described above.

In operation 320, if the path weight for the new path is lower than the path weight for the current path, the subnet manager updates the subnet routing information and disseminates updated routing data. The method then ends.

25 The foregoing embodiments of the invention have been presented for purposes of illustration and description only. They are not intended to be exhaustive or to limit the invention to the forms disclosed. Accordingly, the

scope of the invention is defined by the appended claims, not the preceding disclosure.